Naval Submarine Medical Research Laboratory

NSMRL REPORT 1180 20 AUGUST 92

An Evaluation of the Usability of the MEPSS Prototype Decision Support Program for Abdominal Pain with Two User Populations

by
Elaine F. Chouinard
University of Hartford
with the technical assistance of
Naval Submarine Medical Research Laboratory

Released by
R. G. Walter, CAPT, DC, USN
Commanding Officer
Naval Submarine Medical Research Laboratory

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 074-0188

REPORT SECURITY CLASSIFICATION CLASSIFIED	1b. RESTRICTIVE MARKINGS
SECURITY CLASSIFICATION AUTHORITY	3. DISTRIBUTION/AVAILABILITY OF THE REPORT Approved for public release; distribution unlimited
DECLASSIFICATION/DOWNGRADING SCHEDULE	
PERFORMING ORGANIZATION REPORT NUMBER(S) MRL REPORT 1180	5. MONITORING ORGANIZATION REPORT NUMBER(S) NA
NAME OF PERFORMING ORGANIZATION val Submarine Medical search Laboratory 6b. OFFICE SYMBOL (If Applicable)	7a. NAME OF MONITORING ORGANIZATION Naval Medical Research and Development Command
ADDRESS (City, State, Zip Code) x 900, Naval Submarine Base NLON, oton, CT 06349-5900	7b. ADDRESS (City, State, Zip Code) 8901 Wisconsin Avenue, Bethesda, MD 20889-5606
NAME OF FUNDING/SPONSORING ORGANIZATION me as 7a 8b. OFFICE SYMBOL (If Applicable)	9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER
ADDRESS (City, State, Zip Code) me as 7b	10. SOURCE OF FUNDING NUMBERS PROGRAM ELEMENT NO. 63706N; PROJECT NO. M0095; TASK NO. 005; WORK UNIT ACCESSION NO. DN277023

TITLE (Include Security Classification)
evaluation of the usability of the MEPSS Prototype Decision Support Program r abdominal pain with two user populations

PERSONAL AUTHOR(S)
F. Chouinard

TYPE OF REPORT	13b. TIME COVERED	14. DATE OF REPORT (Year, Month, Day)	15. PAGE COUNT
terim	FROM ____ TO ____	20 Aug 1992	

SUPPLEMENTARY NOTATION

COSATI CODES			18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)
D	GROUP	SUB-GROUP	Computer aided diagnosis; decision making; abdominal pain

ABSTRACT (Continue on reverse if necessary and identify by block number)

fully operational prototype of the MEPSS decision support module for diagnosing abdominal pain was ability tested with representatives from two user populations: Independent Duty Corpsmen and MT-Paramedics. Issues with the user interface that negatively affect the accuracy of the program were ntified. Design recommendations are presented. The two groups differed in their acceptance of the program. arketing strategies to increase acceptance are suggested. The utility of usability testing methodology for ntifying problems with the user interface, and for identifying differences in the needs of target user pulations, was demonstrated.

DISTRIBUTION/AVAILABILITY OF ABSTRACT NCLASSIFIED/UNLIMITED ☐ SAME AS REPORT ☐ DTIC USERS	21. ABSTRACT SECURITY CLASSIFICATION UNCLASSIFIED
NAME OF RESPONSIBLE INDIVIDUAL san D. Monty, Publications Office	22b. TELEPHONE (Include Area Code) (203) 449-3967 22c. OFFICE SYMBOL

An Evaluation of the Usability of the MEPSS Prototype Decision Support Program for Abdominal Pain with Two User Populations

by

Elaine F. Chouinard
University of Hartford
with the technical assistance of
Naval Submarine Medical Research Laboratory

Naval Submarine Medical Research Laboratory
NSMRL Report 1180

Approved and Released by

R. G. Walter
Commanding Officer

Approved for public release; distribution unlimited

SUMMARY PAGE

THE PROBLEM

The MEPSS decision support program needs to be easy to use and accurate in its use. Usability testing is the appropriate methodology for evaluating these issues.

THE FINDINGS

Problems with the user interface which affect the accurate use of the program were identified. Design changes are recommended. Subjective ease of use differed between the two user populations tested.

APPLICATIONS

Accurate and effective use of the MEPSS decision support program will be facilitated by implementation of the Design Recommendations. Analysis of the subjective ease of use findings can suggest strategies for increasing acceptance of the program by the two user groups tested.

ADMINISTRATIVE INFORMATION

This project was conducted under Naval Medical Research and Development Command Work Unit 63706N M0095.005-5010. The views expressed in this report are those of the author and do not reflect the official policy or position of the Department of the Navy, Department of Defense, or the U.S. Government. It was approved for publication on 20 Aug 92. It has been designated as Naval Submarine Medical Research Laboratory Report No. 1180.

ABSTRACT

A fully operational prototype of the MEPSS decision support module for diagnosing abdominal pain was usability tested with representatives from two user populations: Independent Duty Corpsmen and EMT-Paramedics. Issues with the user interface that negatively affect the accuracy of the program were identified. Design recommendations are presented. The two groups differed in their acceptance of the program. Marketing strategies to increase acceptance are suggested. The utility of usability testing methodology for identifying problems with the user interface, and for identifying differences in the needs of target user populations, was demonstrated.

[Blank Page]

An evaluation of the usability of the MEPSS prototype decision support program for abdominal pain with two user populations

by

Elaine F. Chouinard

INTRODUCTION

A computerized medical diagnostic assistance program (MEDIC), to aid Navy Independent Duty Corpsmen in the diagnosis and treatment of illness and injury, has been under development and in limited use for approximately 14 years (Ryack, 1987). The Independent Duty Corpsman (IDC) is the sole medical officer on board a submarine, and, as such, does not have access to medical colleagues for consultation on difficult cases. For this reason, the Navy is developing a library of software tools to aid him, including training, reference, and decision support programs.

A usability study was conducted in the spring of 1990 comparing three different user interfaces to the abdominal pain module of the MEDIC program. A higher user satisfaction rating was associated with visual grouping of related items (a finding consistent with the findings of Tullis, 1980), ordering of items to resemble the typical medical examination, the use of color to highlight information and direct the user, and minimal and consistent steps for data entry. Longer time to complete a screen was associated with a lack of grouping of related items, multiple steps for data entry, a lack of instructions identifying required and optional data entry items, and the exclusive use of upper case text. Confidence in the program-generated diagnoses was found to increase as the user satisfaction rating increased (Chouinard, Ryack, & Stetson, 1991). These general findings, as well as the specific problem areas observed in the video log of the study, were shared with the developers, and a single interface was developed.

The purpose of the current study was two-fold. First, it served as usability study of the new interface. Second, it served to explore the possibility that the program may prove useful to medical personnel outside the Navy community.

METHOD

DESIGN

The primary purpose of the study was to provide additional data to the developers of the program to aid them in producing the most usable product.

Eissenberg and Redish (1989) suggest a two-step approach to usability testing. Initially an exploratory evaluation should be performed to observe the effectiveness of certain features and to identify problem areas. Once the problems have been addressed, and the product has been developed into its proposed final form, a criterion-based pass/fail test is needed.

The usability study performed in 1990, described above, was undertaken to evaluate the strengths of specific features represented in the three interfaces, and to identify problem areas. This served as an exploratory usability test. The current test was a criterion based usability test. Benchmark tasks were identified through interviews with five Subject Matter Experts (two physicians and three Independent Duty Corpsmen) at the Naval

Submarine Medical Research Laboratory. Benchmark values for these tasks were obtained through a paper and pencil survey of the Subject Matter Experts by the experimenter. The survey requested time and error estimates for the specific tasks of logging on, entering one case, changing a previous entry, accessing Help, and retrieving a diagnostic summary of a previous case, for both novice and experienced users of the program. For each task, the highest of the estimates obtained in the survey were used as criteria for the study (as recommended by Rubin, 1990b), except where there was a large gap between the highest estimate obtained and the range into which the other estimates fell. In this case, a value was chosen which made intuitive sense to the experimenter. The complete list of benchmark tasks and values appear in Appendix A. The values for novice users were used for this study, as the study was the first time that any user had seen this version of the program (thus placing all users in the novice category).

Five additional benchmark criteria were added at the request of one of the designers, and these are also listed in Appendix A. Specifically these state that 75% of all users will indicate the top two points of a five point scale when rating the program as "easy to use" and "accurate", and that 99% of all users will report that Help is available. Also, there will be no combination of keystrokes that can place the user outside of the program, and under no circumstances will the program crash.

The secondary purpose of the study was to explore the possibility that the program may prove useful to medical personnel outside the Navy community. To this end, the study was performed with two distinct user groups, both of whom represented potential users of the final product.

One group was the program's intended user group, Navy Independent Duty Corpsmen. The second group was a civilian medical population, EMT-Paramedics.

The similarities of the EMT-Paramedic group with the Independent Duty Corpsmen (IDC) begins with their training. Independent Duty Corpsmen receive 1755 hours of training (W. H. Calver, Detachment Command Master Chief, Naval Undersea Medical Institute, personal communication, July, 1991), and Paramedics receive up to 2000 hours of training (Nevers, 1991). Prior service is required to enter both the Independent Duty Corpsmen and the EMT-Paramedic training programs. Requirements for prior service for Paramedic trainees differ according to region and program (Connecticut Office of Emergency Medical Services, phone contact, July, 1991). The program at the Mattatuck Community College in Waterbury, Connecticut requires successful completion of the course of training for EMT-ambulance (Mattatuck Community College Catalogue, 1989-1990). Generally, Paramedic trainees have served as Emergency Medical Technicians prior to entering the Paramedic training program, and they must pass a pre-test to be accepted (Connecticut Office of Emergency Medical Services, phone contact, July, 1991). Similarly, Navy hospital corpsmen are eligible to apply to the Nuclear Submarine Medical Technician program after they have served six years in the Navy, and must pass eligibility requirements to be accepted (W. H. Calver, personal communication, July, 1991).

The training for both the IDC and EMT-Paramedic includes medical diagnosis and treatment, clinical skills, pharmacology,

(B. Cialfi, President, National Association of Emergency Medical Technicians/Paramedic Society, personal communication, June, 1991).

Another potentially significant difference between the IDC and the Paramedic in the performance of their duties may be the number of actual cases encountered by each. In a study evaluating user acceptance of the original MEDIC module for abdominal pain, four submarines participated for two months. During that time, only four of the total cases reported were abdominal pain (Henderson, et al., 1981). Paramedics respond to an average of 10 calls each work day (Nevers, 1991), some of which, undoubtedly, are abdominal pain. However, records on the frequency of each type of case encountered are not yet maintained in the state of Connecticut (B. Cialfi, personal communication, June 1991).

The infrequency with which a Corpsman is called upon to assess abdominal pain may lead to his becoming rusty in the procedures for collecting data from the patient. User's comments from the first usability study state that one of the advantages of the diagnostic assistance program was that it reminded them to ask questions they may have forgotten, and served as an information gathering aid (Chouinard, et al., 1991). This may be one characteristic of the program that would be perceived as very useful by the IDCs but may not be as highly valued by the Paramedics, whose data collecting skills have not had the opportunity to become rusty, and who have a greater amount of medical backup available to them.

In the literature, the usability of a tool is often defined to be specific to the particular user population and job demands under study. Aspects of usability that are consistently included in most studies of usability are: ease of use, ease of learning (Potosnak, 1988; Rubin, 1990a), and user preference or satisfaction (Rubin, 1990a; Whiteside, Bennet, & Holtzblatt, 1987). For the purposes of this study, ease of use and user satisfaction were included. Outcome variables were performance measures on the experimental tasks, as measures of ease of use, and scores on a user satisfaction questionnaire and responses to interview questions, as measures of user satisfaction.

Specifically, the performance measures were: time and number of errors to log on to the program and choose the abdominal pain module, time to enter one case, time and number of errors to quit a partially completed case, resumption of a case which was quit midway without lost data (successful execution or not successful), time and number of errors to change an entry, time and number of errors to access a particular Help screen, and time and number of errors to retrieve a diagnostic summary of a previous case. User satisfaction was measured by a questionnaire adapted from a questionnaire in the literature, three additional rating items, and a semi-structured interview.

PARTICIPANTS

Ten Navy Independent Duty Corpsmen and one Navy medical student from the Naval Submarine Base in Groton, Connecticut, and eleven EMT-Paramedics (eight with experience as a Paramedic and three recent graduates from the Paramedic training program) from an ambulance service in Waterbury, Connecticut, volunteered to serve as users of the program. All participants were male.

anatomy, physiology, intravenous therapy, emotional crises, and triage (W. H. Calver, personal communication, July, 1991; Mattatuck Community College Catalogue, 1989-1990). Both the IDCs and the Paramedics are required to spend time in clinical rotations. The Paramedics' clinical rotations are in the advanced life support units of a hospital (Mattatuck Community College Catalogue, 1989-1990). The IDCs perform clinical rotations in medical and dental hospitals, as well as in psychiatric units (W. H. Calver, personal communication, July, 1991).

Some differences exist in the training received by each group. The IDCs receive training in radiological fundamentals and controls (fundamentals of radiation, how to measure and control exposure), and radiation administration (the monitoring of sailors' exposure to radiation) (W. H. Calver, personal communication, July, 1991). The Paramedics receive training in laws and regulations related to emergency care (Mattatuck Community College Catalogue, 1989-1990).

Both IDCs and Paramedics are trained to collect the history and symptoms of their patients. The IDCs are trained to do this according to a S.O.A.P. Note (Subjective - patient symptoms, Objective - clinical data, Assessment - corpsman's impressions, Plan - treatment), and the chronological records of care maintained in the patients' health records follow this format (W. H. Calver, personal communication, July, 1991). The Paramedics are trained to collect and record patient data according to a Primary A.B.C. Assessment (Airway, Breathing, Circulation) and a Secondary Head-to-Toe Assessment (W. Campion, Jr., Campion Ambulance Service, personal communication, December, 1991).

Similarities and differences are present in the way that IDCs and Paramedics perform their duties. Both the EMT-Paramedic and Navy IDCs may be called upon to perform triage, which is the prioritizing of a number of casualties of war or other disaster (Dorland's Illustrated Medical Dictionary, 1988). When presented with a patient, the IDC needs to determine if the condition can be successfully treated on board or whether the patient's condition requires an interruption of the mission for medical evacuation. In a similar way, the Paramedic, when presented with a patient, must decide if the patient's condition requires a call to the hospital for medical backup, or if the patient can be safely transported to the hospital in the routine fashion. The decision to evacuate a sailor (MEDEVAC) exposes the submarine's position and can thereby threaten national security (Henderson, Ryack, Moeller, Post, & Robinson, 1981; Ryack, 1987). In addition, it is dangerous to the patient and rescue crew, and is expensive (Henderson, et al., 1981; Ryack, 1987). For these reasons, evacuation of a patient to a land based hospital facility should only be performed when medically necessary. In contrast, for the Paramedic, transportation of a patient to the hospital is the routine.

The IDC is the only medical person on board the submarine and, therefore, does not have access to professional colleagues for consultation on a difficult case. In fact, one of the primary goals for the MEDIC program is to have it act as a source for medical consultation in isolated environments (Ryack, 1987). The EMT-Paramedic has easy access to communication with other medical personnel. In fact, the Paramedic is required to contact the hospital under certain conditions, as dictated by the protocol established by the Medical Control Physician in charge of the Emergency Department of the hospital

MATERIALS

The Program

The MEPSS module for abdominal pain, in development at the Johns Hopkins Applied Physics Lab, was used in this study. Several of the recommendations from the previous usability study by Chouinard, et al., (1991) were incorporated into this version. A major revision is that this version no longer uses dialogue boxes, which were found to be a source of annoyance to the corpsmen users participating in the 1990 study (Chouinard, et al., 1991). Data entry is now performed by two distinct methods. Where continuous values are required, the user types in values directly. For items requiring a choice of a response from a list, the user places the cursor on the desired choice(s) using the arrow keys, "checks" it off using the space bar, then enters the entire screen using the "Enter" key. To indicate that additional choices are available through scrolling, small arrows appear to the right of the lists.

A second revision addresses the order in which the items in the program are presented to the user. The order of the presentation of the items now corresponds exactly to the order in which the items are addressed by the Corpsmen in the actual examination of the patient. This is a revision that was suggested by many users in the previous study (Chouinard, et al., 1991).

Lastly, the range of responses accepted by the respiration and temperature fields has been expanded according to suggestions offerred by the users in the previous study (Chouinard, et al., 1991).

Sample Cases

Three typed sample cases of abdominal pain, randomly chosen from those used for training at the Naval Submarine Medical Research Lab, were used for the study. These cases appear in Appendix B.

User Satisfaction Questionnaire

Users completed a questionnaire composed of items appropriate to this study from a questionnaire developed by Pearson for the measurement of computer user satisfaction (Bailey & Pearson, 1983). The questionnaire had a predictive validity coefficient of .79 and a reliability coefficient of .93. The version used in this study appears in Appendix C.

The following revisions to Pearson's original questionnaire were made for the purposes of this study. Eleven of the original thirty-nine items were retained. For these items, only the original adjective pair scales appropriate to the purposes of the current study were retained. Additional adjective pair scales, designed specifically for the current study, were added. An additional item not included in the original questionnaire, "Method of data entry", was added, with four new adjective pair scales. The "Language" item, with its associated scales, was repeated twice. The first time it refers to the non-medical language used to communicate instructions to the user. The second time, it asks users to rate the medical terminology contained in the program.

The added items and scales were not included in the calculation of the overall satisfaction scores as described below, because to do so may compromise the reliability and validity of the instrument. Responses to these items were tallied for the purpose of collecting descriptive information on users' subjective reactions to the program.

Each item was scored by taking the average of the values marked for its adjective pairs. The overall usability score is a sum of the item scores (Bailey & Pearson, 1983).

Participant Background Questionnaire

Descriptive variables were collected for each participant through the background questionnaire, which appears in Appendix D. Continuous descriptive variables included age, education level, and years of experience. The categorical descriptive variables were gender, professional experience level, experience with computer systems, experience with computer software, prior experience with a medical diagnostic assistance program, and participants' subjective ratings of their own computer skill level. To assess computer and software experience, participants were presented with five items in each category. Participants indicated those items which they had used by placing a check next to them. For both computer and software experience, participants were classified as Novice if no items were checked, Beginner if 1 or 2 items were checked, Intermediate if 3 or 4 items were checked, and Experienced if 5 or more items were checked. Question 1 under Computer Experience asks participants to assess their own level of computer experience as "None", "Minimal", "Moderate", or "Hacker". Participants were asked to indicate prior exposure to a medical diagnostic assistance program by answering "Yes" or "No".

PROCEDURE

All users participated in the study at their work location during their work hours. One Personal Computer work station in a large office at the Naval Submarine Medical Research Laboratory was reserved for the study. Similarly, desk space in a general office area was used at the ambulance service. To simulate the cramped work environments of both user groups, the experimenter used materials already present on the desks and cardboard boxes to create workspaces of approximately 30 inches across by 30 inches deep by 6 feet high. The rooms housing the test areas remained open to traffic by other personnel during the study in both locations. Participants were permitted to be briefly interrupted by their colleagues or superiors during their participation when necessary. In addition, at the ambulance service location, participants were permitted to respond to a call, although every effort was made by the shift supervisor to assign non-participants to calls whenever possible.

The program was installed on a Zenith 286 personal computer system with a Zenith keyboard (industry standard) and color monitor, and this same system was used in both locations. A technician from the Naval Submarine Medical Research Laboratory checked that the program was running in an identical manner in both locations.

Each user was greeted personally by the experimenter. Users were given a packet containing consent form(s), the test cases, and the User Satisfaction and user background questionnaires. The experimenter read from a script to insure that each user was given identical orientation and instructions.

Users were instructed to "think out loud" while they worked, according to the procedure known as "protocol analysis". Protocol analysis encourages user to speak their thoughts, strategies, and questions out loud (Lewis, 1982; Mack, Lewis, & Carroll, 1983; Rubin, 1990b; Shneiderman, 1987). The use of protocol analysis reduces the ambiguity in the interpretation of user actions. For example, if a user's hand hovers momentarily over the keyboard, is this due to confusion over information presented on the screen, inadequate instructions, distraction from other items presented on the screen, or is it an expression of user frustration or boredom? The

user's comments can often indicate the source of these ambiguous actions.

According to the procedures outlined in the IBM report Using the "thinking aloud" method in cognitive interface design (Lewis, 1982) and recommended by Rubin (1990a & 1990b) in his workshop Usability testing of human-computer interfaces and end-user documentation, users were encouraged to continue their verbalizations with the following prompts, " What are you thinking?" or "What is that telling you?" To encourage a user who had encountered difficulty to keep trying, the experimenter said, "What would you do if you were on your submarine/in your ambulance now?"

One video camcorder, located diagonally behind the user, was aimed at the screen and keyboard, and recorded the entire test session. The purpose of the video record was to record user keystrokes, the display on the screen at all times during the users interaction with the program, and user comments.

Users completed each sample case before continuing on to the next case. Each of the three cases was typed on a separate sheet of paper. Users were able to refer to the typed case descriptions as often as they wished during the test situation.

The first task asked of each user was to log on to the program. The experimenter then instructed the user to enter the first case, as the second experimental task. During the pilot study, it was observed that users had difficulty with the method of data entry that required users to "check" their response(s) on a list, using the space bar, prior to entering the responses with the Enter key. (The prompt "Space-check item" appears at the bottom of the screens requiring this method of data entry). Therefore, the following procedure was used. If a user had not used this space-check function by the third screen on which it was required, the Site of Pain at Present screen, the experimenter spoke the following prompt: "Here I would like you to pause and read the bottom line of the screen." If the user did not offer any comment after reading it, or offered a comment which suggested that he did not understand the space-check function, the experimenter asked the following question: "Does the prompt 'Space-Check' suggest anything to you?" If the user then responded in the negative, the experimenter spoke the following explanation, demonstrating the function at the same time: "In this program you indicate your choice by first checking it off, using the space bar, and then entering the entire screen. This enables you to choose more than one response. You can erase a check by hitting the space bar again."

During the second case, the experimenter interrupted each user at the same item and asked them to assume that an emergency requires that they save and quit their work. This constituted the third experimental task.

During the third case, the experimenter interrupted twice, for the fourth and fifth experimental tasks, at the same items for each user. The first time was to instruct the user to return to a specific previous item and change their response to the new one provided by the experimenter. The second interruption was to instruct the user to locate and read the Help documentation on a specific medical term. The last experimental task required the user to retrieve the diagnostic summary from their first case.

While users performed the tasks, the experimenter manually recorded the following information: the tasks and their times, user

keystrokes during the experimental tasks, the number and content of errors, user keystrokes during problems encountered in parts of the program not related to the experimental tasks, and user comments. These data are recommended by Philips and Dumas (1990) and by Rubin (1990a). Following the sessions, the experimenter reviewed the videotapes to add to the record of each session any user actions and comments that may have been missed during the real time recording of the data. In this manner, a complete session log, describing experimental tasks and problem areas, was produced for each user. A sample log appears in Appendix E.

After completion of the tasks, the users completed the questionnaires. During the interview, all users were asked the same set of questions as shown in Appendix F. In addition, users were asked individual questions as they related to their performance of the tasks, and were given time to elaborate, express frustrations, ask questions, and offer suggestions. The background and the purpose of the study was explained, according to the debriefing script which follows the interview questions in Appendix F.

RESULTS

USABILITY RESULTS

Issues Affecting the Accurate Use of The Program

Two problems with the user interface were uncovered by this study which effect the accurate use of the program. Both problems were experienced by both user groups. The problems were with the "Space - check item" method of data entry and with the prompts indicating the continuation of a list of response options.

The "Space - check item" method of data entry.

For items requiring that the user choose a response from a list, the user must highlight his/her choice using the arrow keys, place a check next to the desired response(s) using the space bar, and then enter the entire screen using the Enter key. The prompt "Space-Check Item" appears at the bottom of the screen as the only instructions to the user regarding this data entry procedure. During the study, if a user had not used this space-check function by the third screen requiring it, the experimenter demonstrated the function as described in the Procedure section above.

No users used the space-check function as a result of the instructions appearing on the screen, and six out of 22 users (27%) correctly used the space-check function after it was demonstrated and explained to them. The result of the users' failure to use the space-check function is that the program calculates its diagnosis using incomplete information, a situation which compromises the accuracy of the diagnosis.

More importantly, users were not aware that the diagnosis was calculated from incomplete data (as evidenced by their surprise when examining the View Data screen, and finding the values they thought they had entered missing).

Prompts to scroll for more response options.

When responses are required as choices from a list which extends beyond one screen, the existence of more options is noted by two small up and down arrows which appear to the right of the list. Many users failed to see these arrows and assumed that the first screen display represented all of their response options.

For the questions requiring the user to scroll down to view more response options, eleven users, who indicated that they were looking for a particular response, did not scroll to search for it. These users commented that the program did not offer them enough or adequate response choices. They chose two, less accurate, responses for the one they were looking for (for example, "Right Upper Half" and "Left Upper Half" for "Central"). The result of the users' failure to indicate the most accurate response choice results in the program calculating its diagnosis using incomplete or inaccurate information, a situation which compromises the accuracy of the diagnosis.

Issues With Data Entry

Unacceptable value error messages.
When an unacceptable value was typed into the Height, Temperature and White Blood Count screens, an error message was displayed. When the error message disappeared, the program progressed on to the next screen instead of staying on the current screen for correction. Three users had raised their hands to the keyboard to re-key the information correctly. Upon realizing that the program had moved on, the users did not attempt to go back and re-key the entry into its appropriate field. Instead, they continued with the next question, leaving the previous one unanswered. Three users, not realizing that the program had moved on to the next screen, re-keyed the information into the wrong field (for example, the temperature value was entered in the Respiration field because that is the next screen following Temperature). These situations again result in the program calculating its diagnosis using incomplete or inaccurate information, a situation which compromises the accuracy of the diagnosis.

At the Temperature screen, 10 of the users used the arrow key to place the cursor over the digit to be corrected and typed over that character only. This resulted in the program displaying an error message stating that the inputted value was not in acceptable limits. In addition, when reviewing the case with the View Data screen, only the one character that had been typed over had been accepted as the temperature. This situation also results in the program calculating its diagnosis using incomplete information.

The White Blood Count screen displays an error message stating that the entry is not in the acceptable ranges if it is, in fact, within the acceptable ranges but has been typed in using a comma. This was a source of annoyance for six of the users.

The Date of Birth field requires a leading zero to accept the value. Not all users automatically preceeded the month value with a leading zero, resulting in an error message and the need to re-key the entry.

Field labels.
At the user log on and patient identification screens, two users entered the patient name for user logon, and five users entered their own name, date of birth and Social Security Number for the patient's.

Issues With Navigation
For the problems described below, "unable to complete the task" is defined as entering a part of the program from which it would have been impossible to perform the task without first returning to the Main Menu, or that the participant gave up in his attempt.

Logging on to the abdominal pain module.
Two users were unable to complete this task.

Quitting the program.
Two users were unable to complete this procedure.

Quitting a case.
Seven users, upon completing their first case, experienced difficulty leaving the case and starting the next case. Six users, tried the "Re-enter Diagnosis" and "Change and Rerun" options to start the second case. One user shut off the personal computer system and then turned it on again in order to begin a new case. Only five of these users eventually chose the correct option, "End Session", leaving two who were unable to complete this procedure.

After the Site of Pain at Onset screen during their second case, all users were instructed to "Assume that an emergency has occured that requires that you briefly leave your patient and the computer. You want to be sure that the information you have already entered is not lost. Save this case, then quit the program." It is important to note that the message "Saving encounter..." had been displayed as each user had left their first case using the "End Session" option. Yet, ten out of twenty-two users were hesitant to use the "End Session" option to quit their second case, stating that they were not sure if this option would save it. Six of these users eventually tried "End Session" after exploring other options, leaving four users who were unable to complete this task.

To resume the case, the correct choice from the Main Menu is "Make Medical Diagnosis". After registering the patient, the program then presents the user with the first unanswered question for the case they were working on. Ten of the 22 users initially chose the option "Review Previous Encounters".

Retrieving a previous diagnosis.
From the diagnosis screen of the third case, users were instructed to retrieve the diagnostic summary of the first case. Two users were unable to complete this task.

Patient identification.
When presented with the list of the Most Recent Encounters, users expressed confusion at the program assigned patient ID number. Users chose their Encounter either by the time of day or by returning to the patient registry screen and using the patient's Social Security Number.

Medical Terminology
Users expressed difficulty with the terminology used on the Inspection of Abdomen and Bowel Sounds screens. Nine users commented that the response choices under Inspection of Abdomen relate to "bowel sounds". They were then further confused when they reached the screen labeled Bowel Sounds. Users commented that "Peristalsis" should not be an option on the Inspection of Abdomen screen because peristalsis cannot be seen, and because peristalsis is normal (the Help screen here indicates that the item should be answered with regard to "visible peristalsis", but this is not indicated on the screen itself). One user suggested that the response options should be "palpations, percussions, masses, and distension". At the Bowel sounds screen, another user suggested these response options should be "hyperactive, hypoactive, and normal". At the Inspection of Bowels screen, three users suggested that Inspection of Stools would be a more accurate title. One user observed at the Inspection of Bowels screen, "The Help screen is completely different from what the screen is presenting. Bowel inspection versus bowel habits. It doesn't match."

Comparison with Benchmark Values

Benchmark values were derived from a survey of five Subject Matter Experts (see Appendix A). The values provided for the "novice" user was used to evaluate the data, since each user in the study was working with the program for the first time. The criteria for deciding if the program has passed against a particular benchmark was 75% of users reached the benchmark value, and no user failed in his attempt to perform the task.

Logging on.

The benchmark value obtained for this task was 30 seconds with no more than 2 errors. However, the survey requested Subject Matter Experts to give a time for starting the program, logging on, and then choosing the abdominal pain module. The current configuration of the program requires that the patient be registered, a process which requires the completion of two screens, prior to choosing the abdominal pain module. Therefore, this benchmark value is not a valid comparison for the obtained data.

The average time for users to progress from the Main Menu screen to the first patient registry screen was 23 seconds (15.9 for the Corpsmen and 30.1 for the Paramedics). Two users were unable to complete the task (one Corpsman and one Paramedic).

Two users failed in their attempt to log on to the program, one from each user group. According to this criteria, the program failed against the benchmark criteria for Logging-on.

Entering one full case.

The benchmark value for this task was 30 minutes. The average time to complete the entry of the first case was 13 minutes, 36 seconds (11 minutes, 23 seconds for the Corpsmen and 15 minutes, 36 seconds for the Paramedics). The program passed with both groups on this benchmark task.

Changing an entry.

At the same point in entering their third case, all users were instructed to return to the Temperature screen and change their entry from 97.4 to 97.9. The benchmark value obtained for returning to a previous screen and changing an entry was 90 seconds with no more than 5 "wrong turns".

Three users exceeded the 90 second benchmark value (their times were 115, 119, and 135 seconds). The average time for the users to complete this task was 41.9 seconds (40.6 for the Corpsmen and 43.3 for the Paramedics), well below the benchmark value. No users made more than five "wrong turns". Based on the criteria of 75% of users reaching or exceeding the benchmark value with no failures, the program passed with both groups on this benchmark task.

Accessing help.

All users were instructed to access and read the Help documentation on "distension" at the appropriate time during the inputting of the third case. The benchmark value for accessing Help was 60 seconds, with no more than 5 "wrong turns".

Two users exceeded this value, their times: 73 and 109 seconds. The average time to access Help was 14 seconds (11.5 seconds for the Corpsmen and 17 seconds for the Paramedics). Removing the two outlying values from the calculations produces an average of 6 seconds to complete this task (1.75 for the Corpsmen and 10.7 for the Paramedics). No users made more than five "wrong turns". Given that all users were able to complete this task and more than 75% of the users completed it within the benchmark

time, the program passed with both groups on this benchmark task.

Reviewing a past diagnosis.
At the completion of their third case, users were instructed to review the diagnostic summary for their first case. The benchmark value for calling up a diagnostic summary from the Main Menu was 60 seconds, with no more than five "wrong turns". Five users exceeded the time value and two users, both from the Paramedic group, were unable to complete the task. The average time to complete the task was 56 seconds (51 for the Corpsmen and 65 for the Paramedics). No users made more than five "wrong turns".

The program failed with both groups on this benchmark task. Neither group had 75% of its users able to meet the benchmark value for time to retrieve a previous diagnosis. Only 73% of the Corpsmen and 66% of the Paramedics completed the task within the suggested time period.

Easy to Use, Accuracy, and Help ratings.
In addition to the benchmarks obtained through the survey of Subject Matter Experts, the program was evaluated against five new benchmarks requested by one of its developers. The results follow. "75% of all users will rate the program as 'easy to use', indicating the top two scale points on a five point scale." 65% of all users' ratings fell into the top two scale points. However, a breakdown of the two user groups shows that 89% of the Corpsmen and 36% of the Paramedics rated the program using the top two scale values. The program passed this benchmark value with the Corpsman sample, and failed with the Paramedic sample.

"75% of all users will rate the program as 'accurate', indicating the top two points on a five point scale." 35% of all users' ratings fell into the top two scale points. The benchmark value was not met with either group. "99% of novice users will report that 'Help' is available when desired." 81% of the users responding to this question reported "Help" to be available when needed. One user indicated that "clinical" Help (definitions of medical terminology) was available and "program" Help (navigation instructions) was not. The benchmark value was not met.

"There is no combination of keystrokes that will result in anything but the program either performing one of its functions or displaying an error message." At no time did the program display something that was not a part of the user interface. This benchmark value was attained.

"Under no circumstances will the program 'crash'." There were no instances of the program crashing. This benchmark value was attained.

SUMMARY
In summary, the program failed against four of the eleven benchmark criteria: Logging-on, Retrieving a Previous Diagnosis, Accuracy rating, and Help available rating. The two groups differed on only one of the benchmark criteria, the Easy to Use rating, with the program passing with the Corpsmen and failing with the Paramedics. Table 1 presents a summary of the two groups' performance with the program in comparison to the benchmark values.

Table 1
Comparison with Benchmark Values

Task	Paramedics	Corpsmen
Log On*	Fail	Fail
Enter a Case		
Time	Pass	Pass
Change Entry		
Time	Pass	Pass
Errors	Pass	Pass
Access Help		
Time	Pass	Pass
Errors	Pass	Pass
Retrieve Diag*		
Time	Fail	Fail
Errors	Pass	Pass
Easy Use Rating	Fail	Pass
Accuracy Rating	Fail	Fail
Help Available	Fail	Fail

* One or more users were unable to complete this task

DIFFERENCES BETWEEN THE TWO GROUPS

Performance and preference measures were collected as data in this study. The research hypothesis was that, for each measure, the Corpsmen would perform better, and express a higher preference for the program, than the Paramedics because it was designed with their specific work demands in mind. The differences between the two groups on were analyzed using the t-test for the difference between two means. Given the uni-directional nature of the hypothesis, the one-tail probabilities ($p = .05$) were used. The one-tail t-test tests a null hypothesis that there is no difference between the two groups or that the difference is negative (Guilford, 1978). For those measures where one or more participants were unable to complete the task, the test for the difference between two independently computed proportions (Bruning & Kintz, 1987) was used.

Comparison of Performance Measures for the Two User Groups

The performance variables were in the form of time and error measures. Times to perform the following tasks were taken: log on, enter one case, return to a previous screen and change an entry, access Help, retrieve a previous diagnosis, and time to save and quit a case. Number of errors while performing the following tasks were recorded: log on, return to a previous screen and change an entry, access Help, retrieve a previous diagnosis, and whether or not the case was successfully saved.

Times.

The times for the two groups on the performance variables are presented in Table 2.

The two groups differed significantly ($p < .05$) only on time to enter one case, but not on time to return to a previous screen and change an entry, and time to access Help according to t-tests.

On three of the six tasks, some participants were unable to complete the task when, after a series of "wrong turns", they entered a part of the program from which it would have been impossible to perform the task without abandoning the task and returning to the Main Menu to start over. Two users, one from each group, were unable to complete the task of logging on. Two users from the Paramedic group were unable to complete the task of retrieving a previous diagnosis. Two users from each group were unable to complete the task of quitting the program. For these tasks, it would have been inappropriate to analyze the data using the t-test because the participants' times should be viewed as approaching

Table 2
Mean Times (in seconds) for Paramedics and Corpsmen on Performance Measures

	Condition				
Task	Paramedics	n	Corpsmen	n	value
Log On	30.1	10	15.9	10	z = 0.51
Enter a Case	936.3	11	683.6	10	t = 2.97*
Change an Entry	43.3	10	40.6	11	t = 0.16
Access Help	17.09	10	11.5	11	t = 0.42
Retrieve Diag.	65.2	6	51.2	11	z = 1.03
Quit	57.0	9	44.3	9	z = 0

*p < .05, one-=tailed

infinity. For these tasks, times were categorized as either falling within the benchmark values or exceeding them, and those unable to complete the tasks were placed in the exceeding benchmark category. Using the test for significance of difference between two proportions, a z value "greater than or equal to 1.96 or less than or equal to -1.96 is considered significant at the .05 level" (Bruning & Kintz, 1987, p. 275). None of the \{z\}-scores were significant, thus the two groups did not differ significantly on the proportion exceeding the benchmark values for time to log on, time to retrieve a previous diagnosis and time to quit the program.

Errors.

The error rates for the two groups on the performance variables are presented in Table 3.

It can be argued that an error resulting in the user being unable to complete a task is qualitatively different from a recoverable error. Therefore the error data for the tasks of logging on and retrieving a previous diagnosis were not analyzed because, for both of these tasks, at least one participant was not able to complete the task.

The two groups did not differ significantly on errors in accessing Help.

The difference between the two groups in errors in changing a previous entry ($t(21) = -2.06$) was not in the predicted direction. The use of a one-tail test tests the hypothesis that there is no difference between the groups, or that the difference is negative (Guilford, 1978). That is, "All outcomes not in the

Table 3
Mean Error Rates for Paramedics and Corpsmen on Performance Measures

	Condition				
Task	Paramedics	n	Corpsmen	n	value
Change an Entry	0.00	10	0.64	11	-2.06
Access Help	0.20	10	0.27	11	-0.24

critical region are regarded as generated by chance" (Guilford, 1978, p. 171). For this result, then, the null hypothesis is not rejected, since the resulting t-value did not fall in the critical region of rejection in the predicted direction.

The two user groups also did not differ significantly in whether or not data was lost when quitting midway through a case, as tested by a Chi-square analysis.

Comparison of Preference Measures for the Two User Groups

Users' subjective reactions to the program were gathered a number of ways. All users completed a User Satisfaction questionnaire (see Appendix B), producing a user satisfaction rating for each participant. The item on the questionnaire dealing with confidence in the program's diagnosis was examined separately. In addition, on the back of the questionnaire, users were asked to rate their impressions of the program's accuracy and of the ease with which it can be used, as well as the availability of Help.

Following their sessions, all users were interviewed according to the script which appears in Appendix F.

The additional rating questions were added to the back of the User Satisfaction questionnaire, with the result that some users failed to turn over their questionnaires and, therefore, did not respond to these items. To increase the \{n\}, and thereby strengthen the t-tests on these outcome measures, the means for the respective groups were substituted for these missing values. A review of the interview data showed these users' responses to be typical of their respective groups. The substitution of the means is based on the assumption that their responses to the questionnaire items would have been "typical" as well.

One user failed to complete any part of the questionnaire. In his case, the mean User Satisfaction score and confidence score for his group were also substituted for the missing data. A review of the values obtained for the other outcome variables for this user suggest that his performance and preferences were "typical" for his group, supporting the use of the group means for his missing data.

User Satisfaction Questionnaire results.

The questionnaire data is presented in Table 4.

There was no significant difference between the two user groups on their scores on the User Satisfaction questionnaire and their scores for the individual item regarding their confidence in the diagnosis. The possible range for a single item on the questionnaire is -3 to 3, yielding a range of possible scores on the User Satisfaction questionnaire of -33 to 33. The overall mean score was 12.19, a moderately high score given the possible range. The mean score assigned by the Paramedics was 10.42 and the mean score assigned by the Corpsmen was 13.79.

To calculate a User Satisfaction score to be correlated with the response to the item on confidence in the program-generated diagnosis, the value for the item relating to confidence in the diagnosis was omitted from the calculation of the total User Satisfaction score. This was done so as to avoid a spuriously high correlation, which would occur from having the value for the confidence item included in both of the scores being correlated. The overall mean score assigned to confidence in the program-generated diagnosis was .76, and the means for the Paramedics and Corpsmen were .60

Table 4
Mean User Satisfaction and Rating Scores for Paramedics and Corpsmen

	Condition		
Item	Paramedics	Corpsmen	t-value
User Satisfaction	10.42	13.79	-0.98
Confidence item	0.60	0.91	-0.47
Accuracy Rating	2.87	2.55	1.11
Easy to Use Rating	2.73	2.09	2.00*

Note: Mean group values were substituted for missing values, resulting in an n-value of 11 for both groups.

The item Easy to Use rating was reverse scored. The lower mean for the Corpsmen group indicates a higher rating of the program's ease of use.

*p < .05, one-tailed.

and .91 respectively. The resulting correlation coefficient, between the modified User Satisfaction score and the response to the confidence item, was not significant.

Users indicated their ratings of the accuracy and ease of use of the program by checking a point on a five point scale where 1 is "very", and 5 is "not at all". The mean ratings on how easy the program is to use were: for the Paramedic group, 2.7, and for the Corpsman group, 2.1. Since this item was reverse scored, as described above, the lower mean for the Corpsman group indicates a higher rating of the program's ease of use. This difference was significant ($t(22) = 2.00$, $p < .05$). There was no significant difference between the groups' ratings of the accuracy of the program. A correlation between response to the confidence item and users' rating of the accuracy of the program was significant ($r(17) = +.73$, $p < .01$).

Interview results.

During the interview, the users were asked if they felt that the program would be valuable to them on the job. Users replied with respect to whether they would actually use the program, and the two user groups differed in their responses. The majority of the Corpsmen replied "Yes" (10 out of 11 users). The majority of the Paramedics replied "No" (6 out of 9 users - 2 users were unable to be interviewed due to responding to a call). A test for the significance between two proportions, performed on the proportions in both groups responding "yes", was significant ($z(20) = 32.82$, $p < .05$).

Due to the smaller number of patients encountered by the Corpsmen as compared with the Paramedics, it was predicted that more of the Corpsmen would cite as an advantage of the program the fact that it would serve to remind them of some of the questions they need to ask their patients. This prediction was not supported. When explaining the potential value of the program to them in their day to

day duties, exactly two users from each group stated the program's utility as a "reminder" as a plus.

DISCUSSION

USABILITY RESULTS

Time to Enter the First Case

The program fared very well on time to enter a case when compared to the benchmark value of 30 minutes per case for a"novice" user of the program. In fact, the average times for entering one case (13 minutes, 36 seconds, across both groups) was better than the Subject Matter Experts' suggested time of 15 minutes for "experienced" users. These values are particularly impressive in light of the fact that the users were engaged in "thinking aloud" while entering the case. Researchers who have studied the Thinking Aloud method have found that the process can increase task completion time by as much as 50% (Ericsson & Simon, 1980). It can be argued that the reason for the excellent time on this task is due to the program's adherence to existing standards and suggestions for good user interface design.

Tullis (1980) found that the "chunking" of related information on the screen led to significant increases inperformance speed with a program. This program chunks related information many ways. The patient's identifying information appears grouped together across the top of each screen. Each item and its corresponding field for data entry, or its response choices, are enclosed in a box. All prompts for each screen appear as a list along the bottom of the screen. The program also groups related items temporally, that is, related items follow one another. For example, although only one item appears per screen, all eight questions relating to pain are presented consecutively.

Tullis also found that conventional usage of upper and lower case letters leads to reductions in CRT display reading time by as much as 13% over text presented in all capitals (Tullis, 1983). This program displays conventional usage of upper and lower case letters for screen labels, items, field labels, prompts, menus, error messages, and Help text. Upper case letters are reserved for titles of Help topics and lists of item response choices. Help text is left-margin justified (as opposed to fill-justified), which has also been found to be related to faster reading times (Trollip & Sales, 1986).

Two conventions found by Keister and Gallaway (1983) to be related to improved performance in both speed and accuracy are followed by this program: where multiple data entry fields appear on one screen, the data fields are aligned, and specific screen areas are assigned for the display of error messages, prompts and requests for input.

Consistent with the suggestions of Smith and Mosier (1988), display formats remain consistent from screen to screen, lists are used to display related items (for example, response choices), the cursor appears in a consistent location upon initial display of a data entry field, keystroke actions for cursor positioning differ from those required for data entry, data entry pace is user controlled, and the user can change an entry.

The resemblance of a program to the task for which it is used has also been found to influence performance speed such that increased similarity is related to increased performance speed (Hanson, Payne, Shirley, & Kantowitz, 1981). The version of the abdominal pain module used in this study had been revised so as to present items in the exact order in which they are performed during the Corpsmen's ex-

amination of the patient. Given the results of Hanson et al., it is reasonable to conclude that this revision, which increased the program's resemblance to the task for which it is used, contributed to the speed with which the users entered a case, (and, in particular, for the better times achieved by the Corpsman group).

Accuracy, Data Entry and Navigation Issues

Space-check Function

The finding that none of the users properly used the space-check function to enter their data has serious implications for the accurate use of the program. The result of the users' failure to use the space-check function is that the program calculates its diagnosis using incomplete information, a situation which compromises the accuracy of the diagnosis. More importantly, users were not aware that the diagnosis was calculated from incomplete data (as evidenced by their surprise when examining the View Data screen, and finding the values they thought they had entered missing). This could lead to a false sense of security with the diagnosis offered. One user commented that he did not place any confidence in the program because he had entered the correct data, and the resulting diagnosis was way off base. In fact, he had not entered the data, since he had not used the space-check function. Yet, the program never communicated to him that no data had been accepted for those fields.

This finding points to a need to either rewrite the prompts and instructions on the space-check function in a manner which is clearly understood by the user and emphasizes the importance of following this procedure, or replace the space-check method of entering data with one which is more familiar to the users (by virtue of being used in other common applications). Considering the gravity of the consequences of a false sense of security with the program's diagnosis, it would be prudent to provide additional feedback on what data was accepted by the system to the user prior to presenting the program's diagnosis. A warning could appear whenever a user tries to advance beyond a screen for which no data has been accepted. The View Data screen could automatically appear prior to the Diagnostic summary screen, giving the users the opportunity to check that all their responses had been correctly accepted. Once these changes are made, another usability study would need to be performed to test that the new instructions or data entry method are indeed clearly understood by the user and are accurate in performance during actual usage.

The finding that none of the users properly used the space-check function also suggests an explanation for the low percentage of users assigning a high accuracy rating to the program (the program failed with both groups against the Accuracy benchmark criteria). For 20 of the 22 participants, data is available on the extent of agreement between their diagnosis and the program's diagnosis. Out of 60 opportunities for agreement (3 cases for each of the 20 participants) the user and program's diagnosis agreed only 24 times, for a rate of agreement of 40%. Since, due to the failure to use the space-check function, the program was often calculating its diagnosis with incomplete data, while the participants were reaching their conclusions with complete data, this low rate of agreement is not surprising. Given the low rate of agreement, it follows that the users would rate the program low on accuracy.

Data entry issues.

The confusion of the users at the user logon and patient identification screens, which led to their inputting the wrong information, can perhaps best be avoided by preceding the

name, DOB, and SSN field labels with "User" or "Patient", whichever applies. Suggestions for addressing the remaining data entry problems encountered during the study appear in the Recommendations List below.

Navigation issues.

The fact that two users were unable to complete the task of logging on, two users were unable to complete the task of reviewing a past diagnosis, and four were unable to quit the program while saving the case, indicates that these three functions need to be further studied to identify ways to make them intuitively easy to use. Suggestions appear below.

DESIGN CONCLUSIONS

1. Space - check function - eliminate this function and replace it with the following: hitting the Enter key once chooses an item, hitting the Enter key twice registers an entire screen. This approach lends itself particularly well to carry over when using the program with a mouse in the future, where the procedure would be to click the mouse once on an item to choose it, and click the mouse twice to enter the entire screen.

2. Display an error message when a user tries to move beyond a screen for which no response has been entered. One response option could be to skip that item, thus allowing a user to skip an item.

3. Scrolling for more - make the arrows larger and place them closer to the text, so that they appear in the same visual field as the text.

4. Allow users to correct unacceptable value entries by having the program remain on the screen requiring the correction after displaying the error message stating that the correction is needed.

5. Quit and Resume - a.) Rename the option "End Session" to"Save Case and End Session". b.) Consider renaming the Menu item "Make Medical Diagnosis" to "Make Medical Diagnosis or Continue Previous Session".

6. Appendicitis - Include an item that asks if the patient has had the appendix removed. If the answer is "yes", prevent"appendicitis" from appearing as a probable diagnosis!!!

7. Reviewing a previous diagnostic summary - Eliminate the system-assigned patient ID numbers and use the patient's Social Security Number to identify on the Most Recent Encounters list.

8. In data-entry fields, allow the program to accept correction by both methods: type-over of the incorrect digit/letter, and re-keying of the entire entry.

9. Precede the "Name", "Date of Birth", and "Social Security Number" fields labels with either "User" or "Medical Officer's" where the data requested refers to the Independent Duty Corpsman, and "Patient" where patient data is being requested.

10. For numerical data entry fields, allow the program to accept large values both with and without a comma.

11. Medical Terminology - have the terminology reviewed by a subject matter expert in the medical field, and abide by his/her suggestions on correct terminology.

12. Review the error message and Help files and correct wrap-around problems in the text.

DIFFERENCES BETWEEN THE TWO GROUPS

Time to Enter the First Case

There was a significant difference between the two groups in performance of this task, with the Corpsmen's meantime being less than the Paramedics. Because the Corpsmen and the Paramedics receive similar training in the collection of patient history, signs, and symptoms, it is unlikely that the difference between the two groups can be related to the manner in which they handled the information contained in the sample cases.

The two groups differed in the amount of previous exposure to a medical diagnostic assistance program (x^2 (1, N = 22) = 6, p < .05). Nine out of eleven Corpsmen, as compared to three out of eleven Paramedics, stated that they had used such a program at least once before.

Deck and Sebrechts (1984) describe the Process of learning a new computer program as one of active schema retrieval, testing and correction. Carroll and Mack (1985) point out that the areas of matching between the user's metaphor (knowledge about similar systems that the user brings to the program) and the new program facilitate recognition, and the areas of mismatch between the old and new situations facilitate learning. Those users who had already been exposed to a medical diagnostic assistance program had ideas (schemata, metaphors) on how one should be structured. They were able to quickly transfer this knowledge to the new system, taking advantage of practice effects where the current program functioned in a manner similar to the one to which they had been previously exposed. Where the current program functioned differently, it can be hypothesised that these users were able to learn the new system more quickly because they knew what to look for (just as person who uses word processing programs can look for how a new program "justifies" paragraphs, while a newcomer to word processing must first learn that "justification" of text is possible).

It would appear that the previous exposure to a medical diagnostic assistance program on the part of more of the Corpsmen accounts for the superior performance of this group in time to complete a case. This difference between the two groups can be said to be a difference in amount of new learning required by the groups. Thus, this difference in time to enter a case can be expected to disappear over time, as both groups practice with the program.

User Satisfaction Questionnaires and Ratings

The two groups did not differ significantly on User Satisfaction scores. The individual factors from the original Pearson questionnaire can earn a possible value from -3 to 3. All mean scores for both groups on these factors were positive, with the highest rated factor being Format, "... the design ofthe layout and display of the screen contents." Other factors include Precision, Relevance, Completeness, Non-medical and Medical Language, Instructions, Help, Job Effects, and Confidence, on which the two group mean scores were nearly identical. The two groups differed slightly in their mean ratings on the factors Error Recovery and Overall Understanding. On Error Recovery the Paramedic mean was .03 and the Corpsmen mean was .89, on Overall Understanding the Paramedic mean was .30 and the Corpsmen mean was 1.37. The mean rating assigned by both groups for the factor added to the original Pearson questionnaire, Method of Data Entry, was also positive, the mean for the Paramedics at .80 and the mean for the Corpsmen at .91. These results illustrate an overall positive user satisfaction with the program.

An interesting result of the previous study by Chouinard, et al., (1991), which compared three different interfaces to the abdominal pain module, was a moderately high positive correlation of users' response to the question concerning their confidence inthe system-generated diagnosis with their overall score on the User Satisfaction questionnaire ($r(35) = +.81$, $p < .01$). Users stated they felt "surer" that the program had "understood" their input when the interface was more "usable".

This correlation of User Satisfaction score with confidence in the program's diagnosis, as measured by users' response to that item on the questionnaire, was repeated with the data from the current study. The resulting correlation coefficient was not significant. However, a correlation between ratings of accuracy of the program and confidence in the program's diagnosis was significant at the .01 level. This significant correlation can be interpreted two ways. Either the users formed impressions of the program's accuracy and adjusted their confidence in the diagnosis accordingly, or the users' level of confidence in the program affected their judgment of its accuracy.

Users' level of agreement with the program's diagnosis was coded as the number of cases, out of the three sample cases, in which the user's and the program's diagnosis were the same. Not surprisingly, a positive correlation wasfound for level of agreement with both User Satisfaction scores ($r(18) = +.50$, $p < .05$) and confidence in the diagnosis ($r(18) = +.55$, $p < .05$). It makes sense that the users' confidence in the program's diagnosis would increase or decrease as its agreement with their own diagnosis varied. It also follows that they would be more satisfied with a program that agrees with their professional opinions, and less satisfied with one that disagreed. It is surprising that the correlation between level of agreement and ratings of the program's accuracy did not reach significance because one would expect a program which agreed with the users' professional judgment to be viewed as accurate. This correlation was based on only 14 cases due to missing data. It may be that the results would have been significant if more cases could have been included in the analysis.

The fact that level of agreement with the users' diagnosis may influence users' level of confidence in the program is reflected in comments from the user interviews. Many stated that the level of confidence they would be willing to place in the program would be determined over extended experience with the program. As one user put it, "The more times it agrees with [my diagnosis], the greater confidence I would have."

Many users hesitated to place a value on their level of confidence in the program without information as to the validation of the program itself, that is, what rules it uses to reach its diagnosis, knowledge of the nature and size of the data base used by the program to calculate its diagnosis, and information about who wrote the program (for example, was a physician involved?). This suggests that users also rely heavily on outside sources of validation when determining the amount of confidence they would place in a program.

A study currently in progress by Halgren, Flowera and Cooke (1991) varies the amount and type of information given to the users about an expert system's decision rules. When presenting its choice to the user, the system includes a description of the decision rules it used in one of five formats: natural language and high detail, natural language and low detail, rule-based language and high detail, rule-based language and low detail, or

no explanation. Preliminary results show that subjects were most likely to change their selection of a course of action to coincide with the expert systems' recommended action when information about its decision rules was presented in high detail and in natural language (Halgren, et al., 1991). The authorscaution, however, that such explanations presented in natural language may foster a false sense of security with the system. That is, use of natural language, because it is so easy to understand and "natural", may incline the user to perceive of the system as being more knowledgable than it actually is. The resulting inflated confidence in the system could be dangerous in areas such as medical diagnosis (Halgren, et al., 1991). Perhaps this effect could be countered by also presenting information on the size and demographics of the data base utilized by the system, as well as figures on the accuracy of the program in actual usage. The fact that users in the current study requested such information suggests that it could be useful in determining how much confidence to place in the program's output.

Some users commented that the program did not consider enough possible diagnoses for them to place a high confidence in it. This suggests that face validity may play a role in calibrating a user's trust in a system. That is, the program must appear to be thorough in order to earn the users' confidence.

The above results, taken together with those of the study by Chouinard, et al. (1991), suggest the following factors influence users' levels of confidence in a program's output: perceptions of whether the data had been accepted by the appropriate parts of the program, the extent of its agreement with their professional judgments, outside sources of validation, and face validity.

Another factor that may influence a users trust in the output of a program may be their perceptions of their own computer knowledge and abilities. A correlation between participants' self-rating of computer expertise (from the User Background questionnaire) and confidence in the diagnosis was significant ($(21) = .56$, $p < .05$), supporting this hypothesis.

Interviews

A statistically significant and meaningful difference between the two groups arose in their responses to the question of whether the program would be valuable to them on the job. Ten out of eleven Corpsmen replied in the affirmative. Of these, some felt that the program would be useful in making their original diagnosis, and others stated that its value would be in confirming their own diagnostic impressions. The Corpsman who replied "No" felt the program's picture of the patient to be limited, producing a limited diagnosis. The majority of the Paramedics replied in the negative. Those replying "No" cited their time restrictions as the major impracticality of the program. In the performance of their duties, the Paramedics' aim to limit their patient contact to ten minutes because their goal is to treat and transport the patient to the hospital as quickly as is safely possible. It was also stated that a Paramedic does not need to reach a diagnosis in order to treat a patient's symptoms. The Paramedic users who replied "No" to this question offered that the program would be useful, not as a diagnostic assistance tool, but as a continuing education tool. Many suggested that providing a collection of sample cases to be entered as review would be an interesting way to maintain their professional knowledge. Of the three Paramedics who replied "Yes" to the question regarding value on the job, two did so conditionally. One replied that he felt the

program would be useful in his work only in rural cases where the ambulance is far away from the hospital, and if the program was modified to include expanded treatment suggestions. The other stated that the program would not be useful during patient contact but would be useful "back at the office" by offering suggestions for the "impressions" line of the paperwork. He explained that Paramedics treat symptoms regardless of whether a positive diagnosis is possible. However, the paperwork that is required upon returning to the office requests a diagnostic "impression" be listed.

From a marketing standpoint, a decision to try to sell the program as a diagnostic assistance tool to the Paramedics, based on its acceptance as such by the Corpsmen, would be a flawed decision. However, re-packaging and marketing the program as a training and review tool, including with it a library of sample cases, may result in a product that would sell to the Paramedics. This is only known because the prototype was tested with actual Paramedics. This result supports the practices of user-centered design and of testing a program intended for multiple end-users with representatives from all the target end-user groups.

CONCLUSIONS

Usability has been defined to include a number of components in the literature. The choice of measures for any particular study is made to fit the unique situation. In this study, ease of use, as measured by performance data, and user satisfaction, as measured by preference data, were collected. The research hypothesis was that the two groups would differ in usability with the program, and that, for each measure, the Corpsmen would perform better and express a higher preference for the program than the Paramedics.

The two user groups tested here differed significantly on one out of eleven performance measures: time to enter one case. This difference can be attributed to differences between the groups on previous exposure to a medical diagnostic assistance program. Although this difference between the two groups was statistically significant, it was not enough to separate the groups with regard to the benchmark criteria. Both groups' performance resulted in the program earning a Pass when compared with the benchmark criteria for entering a case. The two groups differed significantly on two preference measures: their ratings of the ease of use of the program and their answer to the interview question regarding their intention to use the program in their actual work. The difference in the Easy to Use rating was enough to separate the groups when compared against the benchmark values, with the program meeting benchmark criteria with the Corpsmen and failing to meet benchmark criteria with the Paramedics. The difference in their stated predictions of whether they would actually use the program was in the hypothesized direction, with a majority of the Corpsmen predicting that they would use the program while the majority of the Paramedics indicated that they would not use it for its intended purpose. These differences are both significant and meaningful, as they impact on user acceptance of the program.

One purpose of usability testing is to predict whether or not the program will actually be utilized by the targeted population(s). This study revealed a difference between the two groups such that the program can be predicted to be accepted and used by the group for which it was originally written and not accepted and used by another user group similar to the first. This finding gives support to the practices of user-centered design and usability

testing with representatives of all targeted user populations. The finding that the method for data entry compromised the accuracy of the program's diagnosis also supports the practice of usability testing programs prior to their release for actual use. In the case of a medical decision support system, compromised accuracy can have serious repercussions.

Areas for Future Research

The issue of trust in an expert system is one that has recently been receiving attention in the literature. Mitta (1991) identifies confidence in a program's solution as one of six variables that enter into her equation for quantifying the usability of an expert system. Sind (1991) discusses the relationship of true versus perceived accuracy in her discussion of the usability of a medical expert system. She states that the ultimate goal of using an expert system is to improve the accuracy of the diagnosis over that which can be obtained by the system or the human diagnostician alone. Ideally, the user should be able to reject the suggestion of the expert system when it is wrong and accept it when it is correct.

The component of true accuracy can be further divided into two parts: the accuracy of the data base and decision rules used to reach the conclusion suggested by the expert system, and the accuracy of the data entry techniques. The first issue is a concern of the writers of the code and the second issue is a concern of the user interface designers.

When data entry methods that lead to inaccuracies in the entry of the data are employed, the resulting conclusion reached by the expert system can be inaccurate. Results from the first usability study performed by Chouinard, etal. (1991) with the abdominal pain module show that data entry methods which appear to the user to be inaccurate can lower the user's confidence in the output of the system. Viewed from the perspective suggested by Sind (1991), this is a positive outcome, as it would tend to lead to the user rejecting the suggestion of the system when the data entry method is questionable. Graver implications, however, are illustrated in the results from the current study where the inaccuracy of the data entry method was not perceived by the user. None of the users in this study correctly used the space-check function, but, more importantly, the users did not realize that they had not used this function correctly. This led to the users erroneously concluding that they had accurately entered the patient data, and that the system had calculated its' output based on complete patient data. From the perspective offered by Sind, this is an unsatisfactory outcome, as it would tend to lead to the user accepting the suggestion of the system when it is inaccurate. This finding underlies the importance of designing user interfaces that are not only accurate and easy to use, but are also obvious in their functioning. It also suggests that when considering the accuracy of a system, both the accuracy of the underlying code and the user interface need to be considered equally.

Another component that enters into the accuracy of the diagnosis reached by the user and system team is the "synergy" of the user-program system itself. Sind proposes many variables that may enter into this synergy, including user preference for the program and the overall usability of the program in a given environment.

The current study, together with the previous study by Chouinard et al. (1991), give support to Sind's proposal that usability, in particular, the component of user preference, do indeed enter into the equation. Specifically, the

results suggest that confidence in the data entry methods, level of agreement with the user's professional judgment, face validity, outside sources of validation in the form of information on the rules and data base used by the system, and users' confidence in their own computer expertise may be important factors in the user's calibration of his/her trust in the system. Future research efforts can vary these components independently, and in combination, in an attempt to describe the relationship of usability with users' calibration of trust in a system. These findings can then be applied to design system interfaces which create a level of confidence that is compatible with the accuracy ofthe system. In this way, the goal of users rejecting the expert system's suggestion when it is wrong and accepting it when it is correct can be reached.

ACKNOWLEDGEMENTS

The author would like to thank Captain Douglas M. Stetson (retired) and James B. Mathews, Ph.D. for their guidance during the planning stages of this project, and William Campion, Jr., Chief Gregory Prunier, and Ms. Ellen Perkins for their help with running the study. Thanks are also extended to Bernard L. Ryack, Ph.D. for making this project possible.

REFERENCES

Bailey, J. E., & Pearson, S. W. (1983). Development of a tool for measuring and analyzing computer user satisfaction. Management Science, 29(5), 530-545.

Bruning, J. L., & Kintz, B. L. (1087). Computational Handbook of Statistics (3rd ed.). Glenview, IL: Scott, Foresman and Company.

Carroll, J. M., & Mack, R. L. (1985). Metaphor, computing systems, and active learning. International Journal of Man-Machine Studies, 22, 39-57.

Chouinard, E. F., Ryack, B. L. & Stetson, D. M. (1991). A comparison of the usability of three versions of a computerized medical diagnostic assistance program for abdominal pain. (Report No. 1172). Groton, CT: Naval Submarine Medical Research Laboratory.

Deck, J. G., & Sebrechts, M. M. (1984). Variations on active learning. Behavior Research Methods, Instruments, and Computers, 2, 238-241.

Dorland, W. A. Newman (Ed.). (1988). Dorland's Illustrated Medical Dictionary (27th ed.). Philadelphia: W. B. Saunders Co.

Eissenberg, T., & Redish, G. (1989, December). First a controlled evaluation, then a test. Software Maintenance News, 7(12), 12+.

Ericsson, K. A., and Simon, H. A. (1980). Verbal reports as data. Psychological Review, 87(3), 215-251.

Guilford, J. P., & Fruchter, B. (1978). Fundamental statistics in psychology and education (6th ed.). New York: McGraw-Hill Book Company.

Halgren, S. L., Flowera, K. A., & Cooke, N. J. (1991). The effect of explanation type on human-expert system interactions. Poster presented at the 35th Annual Meeting of the Human Factors Society, San Francisco.

Hanson, R. H., Payne, D. G., Shirley, R. J., & Kantowitz, B. H. (1981). Process control

simulation research in monitoring analog and digital displays. Proceedings of the Human Factors 25th Annual Meeting, 154-158.

Henderson, J. V., Ryack, B. L., Moeller, G., Post, R., & Robinson, K. D. (1981). Use of a computer-aided diagnosis system aboard patrolling FBM submarines: Initial at-sea trials. (Report No. 938). Groton, CT: Naval Submarine Medical Research Laboratory.

Keister, R. S. & Gallaway, G. R. (1983). Making software user friendly: An assessment of data entry performance. Proceedings of the Human Factors Society 27th Annual Meeting, 1031-1034.

Lewis, C. (1982). Using the "Thinking-aloud" method in cognitive interface design. (Report No. RC 9265 #40713). Yorktown Heights, NY: IBM Thomas J. Watson Research Center.

Mack, R. L., Lewis, C. H., & Carroll, J. M. (1983, July). Learning to use word processors: Problems and prospects. ACM Transactions on Office Information Systems, 1(3), 254-271.

Mattatuck Community College, 1989-1990 Catalogue. Waterbury, CT.

Mitta, D. A. (1991). A methodology for quantifying expert system usability. Human Factors, 33(2), 233-245.

Nevers, R. (1991, May 13). Long days, low pay and excitement. Waterbury Republican-American, p. 2D.

Philips, B. H., & Dumas, J. S. (1990). Usability testing: Identifying functional requirements for data logging software. Proceedings of the Human Factors Society 34th Annual Meeting, 295-299.

Potosnak, K. (1988, March). Setting objectives for measurably better software. IEEE Software, 89-90.

Rubin, J. (1990a). Usability Testing of Human-Computer Interfaces and End-User Documentation. Holmdel, N.J.: author.

Rubin, J. (1990b, October). Usability Testing of Human-Computer Interfaces and End-User Documentation. Workshop presented at the 34th Annual Meeting of the Human Factors Society.

Ryack, B. L. (1987). A Computer-Based Diagnostic/Information Patient Management System for Isolated Environments. MEDIC Ten Years Later. (Report No. 1089). Groton, CT: Naval Submarine Medical Research Laboratory.

Shneiderman, B. (1987). Designing the User Interface: Strategies for Effective Human-Computer Interaction. Reading Massachusetts: Addison-Wesley Publishing Company.

Sind, P. M. (1991). Evaluating the effectiveness and usability of MEPSS. Unpublished manuscript, Naval Submarine Medical Research Laboratory, Groton, CT.

Smith, S. L. & Mosier, J. N. (1988). Guidelines for designing user interface software. MTR 10090, The MITRE Corporation, Bedford, MA.

Trollip, S. R., & Sales, G. (1986). Readability of computer-generated fill-justified text. Human Factors, 28, 159-167.

Tullis, T. S. (1980). Human performance evaluation of graphic and textual CRT dis-

plays of diagnostic data. Proceedings of the Human Factors Society (1980), 310-311.

Tullis, T. S. (1983). The formatting of alpha-numeric displays: A review and analysis. Human Factors, 25, 657-682.

Whiteside, J., Bennett, J., & Holtzblatt, K. (1987, November). Usability engineering: our experience and evolution. (Report No. DEC-TR 547). Digital

.[Blank Page]

APPENDIX A

BENCHMARK TASKS

Task Description	Criteria Novice	Expert
log onto program, choose abdominal pain module:	30 sec	10 sec
# errors	2	0
enter one case:	30 min	15 min
# undetected errors:	5	3
# detected & corrected errors:	10	5
return to previous item and change entry:	90 sec	45 sec
# "wrong turns":	5	2
access Help on specific item:	60 sec	10 sec
# "wrong turns":	5	2
from main menu, call up diagnostic summary	60 sec	30 sec
# "wrong turns":	5	2

75% of all users will rate the program as "easy to use," indicating the top two scale points on a five point scale.

75% of all users will rate the program as "accurate," indicating the top two scale points on a fi ve point scale.

99% of novice users will report that Help is available when desired.

There is no combination of keystrokes that will result in anything but the program either performing one of its functions or displaying an error message.

Under no circumstances will the program "crash."

[Blank Page]

APPENDIX B

SAMPLE CASES

Case 1
Date of Birth: 5/1/52

History

This patient is a 39 year old male who presents with pain in his abdomen which began in the lower half and the central part of his belly and is now located in the lower half and seems to come and go. The pain began less than 12 hours ago and is a really intense pain and it seems to be getting worse. Movement seems to made the pain worse and applying heat to the area of pain seems to help a little. He has felt sick to his stomach all day and has not been vomiting. He states that he has not felt like eating today because of his discomfort. He hasn't noticed any change in the color of his skin or eyes recently. His bowels have been relatively normal and he complains of having to urinate more often than usual. He has been bothered by minor G-I upset from time to time and he cannot recall ever having a pain like this before. An appendectomy was performed when he was very young and he doesn't remember any other hospitalization. The patient denies a history of G-I illnesses and is not taking any medication for this pain.

Physical

On physical examination of your patient, he is noted to have a temperature of 100.2, pulse 74, blood pressure 122/80, and his white blood cell count is 8.800.

Your examination reveals a patient who is in obvious distress from his pain and who appears pale. Inspection of the abdomen reveals no abnormalities. No bowel sounds could be appreciated. A surgical scar is present in the midline and there is a generalized swelling of the entire abdomen. The patient reflexively tenses his abdominal muscles during palpation and the abdomen is soft during palpation. There are no masses and Murphy's sign is not present. Tenderness is noted in the middle of the abdomen and rebound tenderness is appreciated. Rectal examination reveals generalized tenderness.

Case 2
Date of Birth: 2/14/42

History

This patient is a 49 year old male who presents with pain in his abdomen which began in the right upper quadrant and is now located in the upper half and seems to be fairly constant. The pain began less than 12 hours ago and is a really intense pain and it seems to be about the same as when it first began. Breathing seems to make the pain worse and vomiting relieves the pain a little. He has felt sick to his stomach all day and has not been vomiting. He states that he has not felt like eating today because of his discomfort. He hasn't noticed any change in the color of his skin or eyes recently. He has had some diarrhea recently and his urinary habits have been normal. He has been bothered by minor G-I upset form time to time and he relates an episode of pain very similar to this a couple of months ago. Repair of a hernia has been his only hospitalization. The patient denies a history of G-I illnesses and is not taking any medication for his pain.

Physical

On physical examination of your patient, he is noted to have a temperature of 101.1, pulse 110, blood pressure 144/94, and his white blood cell count is 10,800.

Your examinations reveals a patient who is in obvious distress from his pain and who appears pale. The patient experienced difficulty in raising his belly to touch your hand when requested to during the abdominal inspection. No bowel sounds could be appreciated. There are no surgical scars on the abdomen and there is a generalized swelling of the entire abdomen. The patient reflexively tenses his abdominal muscles during palpation and there is some residual muscle spasm throughout the examination. There is a mass appreciated centrally and Murphy's sign is present. Tenderness is noted in the right upper quadrant and rebound tenderness is appreciated. The rectal examination is non-revealing.

Case 3
Date of Birth: 4/4/66

History

This patient is a 25 year old male who presents with pain in his abdomen which began in the right half and is now located in the right half and seems to be fairly constant. The pain began less than 12 hours ago and is a really intense pain and it seems to be getting worse. Movement seems to make the pain worse and nothing he does makes the pain any better. He has felt sick to his stomach all day and has not been vomiting. He states that he still feels like eating in spite of the discomfort. He hasn't noticed any change in the color of his skin or eyes recently. His bowels have been relatively normal and he has noticed a red tint to his urine recently. There is no history of previous G-I upset and he relates an episode of pain very similar to this a couple of months ago. An appendectomy was performed when he was very young and he doesn't remember any other hospitalization. The patient denies a history of G-I illnesses and is now taking aspirin and Maalox for his pain.

Physical

On physical examination of your patient, he is noted to have a temperature of 97.4, pulse 74, blood pressure 86/62, and his white blood cell count is 6,800.

Your examination reveals a patient who is in obvious distress from his pain and who appears pale. Inspection of the abdomen reveals no abnormalities. Bowel sounds are normal. A surgical scar is present in the midline and there is no apparent distension. The patient reflexively tenses his abdominal muscles during palpation and there is some residual muscle spasm throughout the examination. There are no masses and Murphy's sign is not present. Tenderness is noted in the right flank area and rebound tenderness is not appreciated. The rectal examination in non-revealing.

[Blank Page]

APPENDIX C

USER SATISFACTION QUESTIONNAIRE

Participant Number _____ Version _____

The purpose of this questionnaire is to measure how you feel about the software program for abdominal pain that you just used. The results will be used to identify ways to improve the program (not to identify dissatisfied users)!

The following pages contain items relating to the software program. Following each item are adjectives for you to use to rate that item. The scale positions are defined as follows:

EXAMPLE

Flexibility of the System: The capacity of the system to change or adjust in response to new circumstances, conditions. or demands.

Adjective A :_:_X_:_:_:_:_: Adjective B
(1) (2) (3) (4) (5) (6) (7)

(1)	extremely A	(5)	slightly B
(2)	quite A	(6)	quite B
(3)	slightly A	(7)	extremely B
(4)	neither A nor B; equally A or B		

According to the response in the above example, the rater felt the system's flexibility was quite "A".

INSTRUCTIONS

1. Respond by placing an "X" on each scale in the position that best describes your evaluation of the factor.
2. Mark only one response for every scale; do not omit any.
3. Place the "X" in a space, not between spaces.

   Correct: :_:X:_:

4. Rely on your first impressions.

1. **Precision:** How closely did the images on the screen match what you expected to see?

high precision	:_:_:_:_:_:_:_:	**low precision**
doubtful (precision)	:_:_:_:_:_:_:_:	**definite** (precision)

2. **Relevance:** How closely did the visual appearance of the output information match your needs? (Output information is anything and everything that appeared on the screen as you used the program.)

useful	:_:_:_:_:_:_:_:	**useless**
relevant	:_:_:_:_:_:_:_:	**irrelevant**

3. **Completeness:** Did the screens provide you with enough information to use the program?

sufficient	:_:_:_:_:_:_:_:	**insufficient**
adequate	:_:_:_:_:_:_:_:	**inadequate**

4. **Format of output:** Please rate the design of the layout and display of the screen contents.

simple	:_:_:_:_:_:_:_:	**complex**
readable	:_:_:_:_:_:_:_:	**unreadable**
useful	:_:_:_:_:_:_:_:	**distracting**
organized	:_:_:_:_:_:_:_:	**cluttered**
professional	:_:_:_:_:_:_:_:	**unprofessional**
easy to understand	:_:_:_:_:_:_:_:	**difficult to understand**

5. **Language:** Please rate the (non-medical) vocabulary used to communicate with the computer program.

complex	:_:_:_:_:_:_:_:	**simple**
powerful	:_:_:_:_:_:_:_:	**weak**
easy-to-use	:_:_:_:_:_:_:_:	**hard-to-use**

6. **Language:** Please rate the (medical) vocabulary used in the computer program.

complex	:_:_:_:_:_:_:_:	**simple**
powerful	:_:_:_:_:_:_:_:	**weak**
easy-to-use	:_:_:_:_:_:_:_:	**hard-to-use**

7. **Error Recovery:** Please rate the methods for making corrections and reentering corrected inputs. How do they compare with your expectations of what a program should provide?

simple	:_:_:_:_:_:_:_:	**complex**
fast	:_:_:_:_:_:_:_:	**slow**
superior	:_:_:_:_:_:_:_:	**inferior**
complete	:_:_:_:_:_:_:_:	**incomplete**
easy to access	:_:_:_:_:_:_:_:	**difficult to find**
easy to understand	:_:_:_:_:_:_:_:	**difficult to understand**
easy to use	:_:_:_:_:_:_:_:	**difficult to use**

8. **Documentation:** Please rate the on-line instructions for the use of the program that appeared on the screen.

clear	:_:_:_:_:_:_:_:	**hazy**
unavailable	:_:_:_:_:_:_:_:	**available**
complete	:_:_:_:_:_:_:_:	**incomplete**
current	:_:_:_:_:_:_:_:	**obsolete**

9. **Documentation:** Please rate the on-line help that appeared on the screen when requested by you.

clear	:_:_:_:_:_:_:_:	**hazy**
unavailable	:_:_:_:_:_:_:_:	**available**
complete	:_:_:_:_:_:_:_:	**incomplete**
current	:_:_:_:_:_:_:_:	**obsolete**
relevant to task at hand	:_:_:_:_:_:_:_:	**useless**
easy to access	:_:_:_:_:_:_:_:	**difficult to find**

10. **Method of Data Entry:** Please rate the method used to key in your history and physical exam data in this program version.

tedious	:_:_:_:_:_:_:_:	**speedy**
simple	:_:_:_:_:_:_:_:	**complex**
easy to use	:_:_:_:_:_:_:_:	**confusing to use**
error prone	:_:_:_:_:_:_:_:	**efficient**

11. **Understanding of System:** Please rate the degree of understanding that you experienced while using the computer-based MEDIC decision support system.

insufficient	:_:_:_:_:_:_:_:	**sufficient**
complete	:_:_:_:_:_:_:_:	**incomplete**
comfortable to use	:_:_:_:_:_:_:_:	**intimidating**
felt in control	:_:_:_:_:_:_:_:	**felt helpless**

12. **Job Effects:** Please rate the changes in job freedom and job performance that you think may result from regular use of the decision support system.

inhibiting	:_:_:_:_:_:_:_:	**liberating**
significant	:_:_:_:_:_:_:_:	**insignificant**
good	:_:_:_:_:_:_:_:	**bad**
valuable	:_:_:_:_:_:_:_:	**worthless**

13. **Confidence in the System:** How much confidence do you have that the diagnosis provided by the program can be helpful to you in formulating your diagnosis?

high	:_:_:_:_:_:_:_:	**low**

Why?

	Very				Not at all
1. This program is easy to use.	1	2	3	4	5
2. This program is accurate.	1	2	3	4	5

3. Did you find HELP documentation to be available when you needed it?

YES NO

APPENDIX D

PARTICIPANT BACKGROUND QUESTIONNAIRE

Participant Number: _____ Date: _____
Date of Participation:_____

Sex: M___ F ___
Date of Birth: _____

Experience/Education

1.) Length of time serving as a corpsman/paramedic:

_____

If Corpsmen: How much time was on a submarine?

_____

If Corpsmen: How much time was on a boat?

_____

If Paramedic: How much time was on an ambulance?

_____

If Paramedic: How much time was "other" (fire department, for example)?

_____

2.) When did you graduate from ID Corpsmen or EMT Paramedic school?

_____

3.) What is your general education level? (Please circle the highest level completed)

High school:	Highest Degree earned:	
	1	High School graduate
	2	Associate
	3	Bachelor
	4	Master
		PhD

College year completed:

1
2
3
4
5

6

Other: _____

Computer experience

1.) How would you describe your level of experience with computers? (Please circle one option)

None

Minimal

Moderate

Hacker

2.) Which types of computers listed below have you used? (Please check all that apply)

___ IBM/IBM clone ___ Mainframe

___ Apple/Macintosh ___ Automatic Teller Machine (ATM)

___ Other (please specify): _____

3.) Which types of programs listed below have you used? (Please check all that apply)

___ word processing ___ spreadsheet

___ computer games ___ Windows applications

___ Other (please specify): _____

4.) Have you ever used a computerized medical diagnostic assistance program?

Yes___ No ___

If Yes, please specify: _____

[Blank Page]

APPENDIX E

SAMPLE SESSION LOG

Participant: 12 **Date:** 7/5/91 Time: 3:15p

Participant Actions	Participant Comments
Task: Log On. Elapsed time: 14 seconds.	
At patient register screen: P types in own SSN enters patient's name	"My last name?"
At Temperature screen: types in value <E>	"Do I have to put in Farenheit? We'll find out!" "No."
Space-Check items.	
Does not use space bar.	
Investigator prompt, re: space-check.	
Investigator question, re: space-check.	"No idea."
Investigator explanation P uses space-check"	"Thank-you." I don't see how that makes a difference."
At Aggravating factors screen: tries to type in a response uses space to check "Other" tries to type in a response	"Why do you have "Other" if you can't type in what "Other" is? Doesn't make sense.
F1 reads Help documentation	"I'm reading Help to try to find out how to enter other aggravating factors and I can't find out how. I guess I'll just go on."

Esc
reads Options menu
goes on to next item

At Relieving Factors screen:

"Here we go again, it's the same thing."

At Appetite screen:

"Suggests a chronic thing, cancer, so I'd be inclined to answer 'No', even though the case says 'Yes'."

P chooses "yes"

P interrupted to go out on call.
7:20 - session resumed.

P realizes that his previous responses were not entered due to not using the space bar. P returns to start of case and begins again.

"If I only have one [response], I don't have to check it?"

Investigator response:
"The way it operates now, you do."

"If it's already checked - just hit Return?"

Investigator response: "Yes"

At Aggravating Factors screen:

"I never did figure out this "Other". I want to put palpatations here. If you hit Help, it just defines things. I give up! I gave up before, I'm giving up again."

Investigator explains that it is not possible to elaborate on what the "Other" is.

At Inspection of Bowels screen:	"If there's only one possible choice, do I still have to check it?"
At Previous Illness screen:	"If 'Yes', where do you get to indicate what it is?"
F1 reads Help documentation	"... 'add details to the back of the'... page down... 'data sheet'. We don't have a data sheet!" (Help documentation instructs user to indicate what the illness was on the data sheet)
At Inspection of Abdomen screen:	"This is very terrible [inadequate choices]. Normal? Peristalsis? Decreased? actually bowel sounds."
At Bowel Sounds screen:	"Here are bowel sounds. That's wierd. Now we have to go back and see what they mean [at Inspection of Abdomen screen]. What do you suppose they mean? Esc, back, esc, back, esc, back (P is narrating his actions) Inspect means to look. How can you look to see peristalsis? That's a really dumb question."
At Rebound Tenderness screen:	"Middle of abdomen... Where do you think that is? I guess what we can do is say... I wonder, if there's no... I'll give... a check there and a check there."
P chooses Upper Half and Lower Half	
At enter user diagnosis screen: P verbally considers each diagnosis in turn, reviewing the pertinent signs and symptoms out loud. P moves cursor to choice and hits space bar, no system response	"Oh, it doesn't work here."

P reads bottom line.
P hits <E>.

Leaving case 1:
<E>
P reads bottom line
Esc

"Change data and rerun?
No. Quit? No."

F1 (Help)
Esc
chooses "End Session"

"This isn't helping."

"I hope I'm not ruining anything!"

starting case 2:
at patient register screen:
P tries 3 times to enter patient date of birth as 2/14/42, system does not accept, re: requires 0 before 2.

"What am I doing wrong here? Oh, I see, I have to type 02."

Task: Quit. Elapsed time: 1 minute, 8 seconds.

Esc
reads Options menu

"End Session? I don't know how to save it."

F1
reads Help documentation

"I hope this is not a major emergency!" (P is referring to instructions for this task.."Assume an emergency has called you away..." and the time it is taking him to discover how to quit the program while saving the case)

P reads in Help documentation where it says all data is saved with the End Session command
chooses "End Session"

Task: Resume. Data lost? No.

P follows correct sequence for resumption of case 2, is presented with first unanswered screen

"We already did this - do I have to do it again?"

At Inspection of Abdomen screen:	"I still don't know what it means."
At Tenderness screen chooses Right Upper Quadrant and Left Upper Quadrant, changes to RUQ only	"..got pain everywhere, 'generalized' would be a nice one [option] to have."
Leaving case 2: PgUp PgDn arrow keys	"How did I get out of this last time? Do you have any idea?"
	"Alright - ask for Help!"
F1 (help) Esc reads "Change and Rerun"	"No."
returns to diagnosis screen	"That's what we should do is just quit because it will save it and quit - just like last time."
Esc Esc	"It says Esc for quit."
P explores on-line references.	"That's pretty cool!" (P comments that he likes the content of the on-line references.)

Task: Change Entry. Elapsed time: 7 seconds

Esc chooses "Previous Question" until at Temperature screen types over entire entry <E> repeatedly to return to current screen.	"I did that pretty fast, didn't I?"
At Site of Pain at Onset screen: P using down arrow key to go to bottom of list, holds key too long and other options scroll into view.	"Ooooh!!! Look at this, look at this! Those tricksters! I was looking for lower abdominal pain before. Look at this. You hid it!!! How many people found

	that? I get an A for finding that one! That could be why you're getting the wrong diagnosis!!"
At Relieving Factors screen: P holds down arrow key for an extended period of time to check for other "hidden" options	
At Medication screen: Esc	"Don't you want to know what they are?"
reads Options menu	"Looking for 'add data'." "Continue current.. it was
chooses "Continue Current Question"	right under my nose." "No, that's not it."
At Site of Tenderness screen:	P felt he had done this item before
F1 reads Help documentation	"Their 'rigidity' is what my 'guarding' is."

Task: Access Help. Elapsed time: 24 seconds
Esc
F1
Esc, Esc
F1

Task: Review Diagnosis. Elapsed time: 45 seconds

Esc looking for "Review Previous.."
chooses "End Session"
at Main menu, chooses "Review Previous Encounters"
chooses "Another Patient"
enters SSN, name
chooses patient
accepts patient
chooses "View on Screen"

APPENDIX F

DEBRIEFING SCRIPT

1. Do you think this program would be valuable to you on the job? Why or why not?

2. The program was designed to serve as a medical consultation for difficult cases, not to replace your professional judgements. How much confidence would you place in the diagnosis offerred by the program, if you were consulting it as a second opinion on a real case?

3. This study is what is called a usability study. It is designed to test not only the functionality of a program, but also how easy it is to learn, and how easy or "normal" it feels to use.

3a. Can you offer any comments on how it felt to use the program?

3b. Was it confusing or intimidating, or did it feel natural (ask for specific examples)?

4. Did you run into any particularly frustrating "gliches"?

5. Did the use of color in the program enhance its use, distract you, or were its effects neutral?

6. In your opinion, did the program show the information on the screen fast enough as you moved from one item to another?

7. Do you have any additional comments or suggestions?

I'd like to let you know how the information we collected today will be used. The program we evaluated is part of a library of programs being developed by the Navy for use by corpsmen on board submarines. The developers feel the library may also prove valuable to medical personnel in other somewhat isolated environments. So, in this study, I am asking both corpsmen and paramedics to evaluate the program. The results of the study will be given to the Navy and will be used to improve the operation of the program. The results will also be available to anyone else who is interested, probably by the late fall. (I will take the names and addresses of anyone who would like to receive a summary of the results). I designed, and am running the study, as my thesis to meet the requirements for my Master's degree.

Thank you for your participation in the study.

[Blank Page]